Detecting Textual Saliency in Privacy Policy

DS 5500
FALL 2021

Phase-2 Presentation

Kaushik Nishtala ■ Shubhanshu Gupta

Goal

Provide key information to end-users using Machine Learning and Natural Language Processing

INFORM
Module

- Visualization
- Annotating policies

QUERY
Module

- Enable users to ask policy-related

Interactive

Web Framework

Segment Classification
(INFORM Module)

- ★ Improve baseline CNN performance with Domain-specific word-embeddings .
- ★ Domain specific word-embeddings produced by fastText are non-contextual.

BERT

- ★ Takes context into account.
- ★ Employs numerous layers of transformer encoders.
- ★ Use pre-trained BERT on vast amounts of unlabeled data, followed by fine-tuning on specific labeled data to solve our downstream tasks.

Segment Classification
(BERT)

- Vocabulary size of 30,522 and uses WordPiece tokenization.
- Helps capture character or subword level information - detect similar prefixes/suffixes/common-roots among tokens.
- Pretrained using two unsupervised tasks - MLM and NSP.
  - MLM: Predict 15% of the randomly "corrupted" or "masked" tokens in a sentence.
  - NSP: Recognize when a pair of sentences appear (or do not appear)
- This way, the model learns an internal representation of the English language that can then be used to extract features useful for downstream tasks.
- Experimented with both case and uncased model types.

Training BERT classifiers

- Modify the head of the architecture by adding a linear layer with dimensions similar to the number of categories in the dataset (12).
- Convert to multilevel classification - used sigmoid instead of softmax to get the probabilities and optimized the binary cross-entropy loss.
- Fine-tuning BERT lasted for 31 hours for two epochs on a single CUDA-backed GPU. Training on the classification task took only a few hours, for five epochs.

Vanilla BERT-base (uncased)

PrivaSeer corpus of 1M privacy policies

Domain-specific RoBERTa-base (cased)

12 encoder layers
Hidden state size of 768
12 attention heads
108M parameters

Convolutional Neural Networks

(Classification)

Category	Majority-vote based dataset		Union-based dataset	
	CNN (Pre-trained)	CNN (Domain-specific)	CNN (Pre-trained)	CNN (Domain-specific)
	F1	F1	F1	F1
First Party Collection/Use	0.82	0.83	0.80	0.82
Third Party Sharing/Collection	0.76	0.81	0.78	0.79
User Access, Edit & Deletion	0.45	0.65	0.60	0.62
Data Retention	0.15	0.37	0.3	0.48
Data Security	0.81	0.82	0.71	0.73
International and Specific Audiences	0.83	0.83	0.92	0.92
Do Not Track	0.88	1.00	0.33	0.61
Policy Change	0.83	0.89	0.63	0.76
User Choice & Control	0.69	0.75	0.65	0.65
Introductory/Generic	0.63	0.75	0.63	0.65
Practice Not Covered	0.09	0.15	0.31	0.39
Privacy Contact Information	0.64	0.88	0.76	0.75
Micro Average	**0.741**	**0.791**	**0.703**	**0.719**
Macro Average	**0.634**	**0.713**	**0.619**	**0.662**
Weighted Average	**0.72**	**0.746**	**0.685**	**0.701**

Table 1: F1 scores on the test set for CNN (w/ pre-trained word embeddings) and CNN (w/ Domain-specific word embeddings) on both datasets (in decimals).

Bidirectional Encoder Representations from Transformers (BERT) (Classification)

Category	Majority-vote based dataset							
	Vanilla BERT				Domain-specific RoBERTa			
	Precision	Recall	F1	Support	Precision	Recall	F1	Support
Data Retention	0.00	0.00	0.00	11	0.75	0.27	0.40	11
Data Security	0.89	0.75	0.81	32	0.83	0.78	0.81	32
Do Not Track	0.00	0.00	0.00	4	1.00	0.75	0.86	4
First Party Collection/Use	0.90	0.87	0.89	181	0.88	0.93	0.91	181
International and Specific Audiences	0.94	0.76	0.84	45	0.91	0.91	0.91	45
Introductory/Generic	0.91	0.69	0.78	58	0.93	0.64	0.76	58
Policy Change	1.00	0.94	0.97	18	0.72	1.00	0.84	18
Practice Not Covered	0.00	0.00	0.00	19	0.82	0.47	0.60	19
Privacy Contact Information	0.85	0.74	0.79	31	0.92	0.71	0.80	31
Third Party Sharing/Collection	0.87	0.87	0.87	142	0.90	0.94	0.92	142
User Access, Edit and Deletion	0.94	0.73	0.82	22	0.95	0.86	0.90	22
User Choice/Control	0.83	0.66	0.74	53	0.82	0.79	0.81	53
Micro Average	**0.89**	**0.76**	**0.82**	**616**	**0.88**	**0.84**	**0.86**	**616**
Macro Average	**0.68**	**0.58**	**0.63**	**616**	**0.87**	**0.75**	**0.79**	**616**
Weighted Average	**0.84**	**0.76**	**0.80**	**616**	**0.88**	**0.84**	**0.85**	**616**

Table 2: F1 scores on the test set for BERT (pre-trained) and RoBERTa (Domain-specific) on the majority-vote dataset (in decimals).

Comparing to previous benchmark results

1-5% improvement over previous results

	F1-scores on the test split (majority-vote dataset)		
Category	Polisis (Harkous et al.)	Establishing strong baseline (Nejad et al.)	Our work
First Partly Collection/Use	0.79	0.91	0.91
Third Party Sharing/Collection	0.79	0.90	0.92
User Access, Edit & Deletion	0.80	0.73	0.90
Data Retention	0.71	0.56	0.40
Data Security	0.85	0.80	0.81
International and Specific Audiences	0.95	0.83	0.91
Do Not Track	0.95	1.00	0.86
Policy Change	0.88	0.90	0.84
User Choice & Control	0.74	0.81	0.81
Introductory/Generic	0.70	0.79	0.76
Practice Not Covered	0.70	0.35	0.60
Privacy Contact Information	0.87	0.78	0.80
Micro Average	0.81	0.85	0.86
Macro Average	-	0.79	0.79
Weighted Average	-	-	0.85

What worked for us?

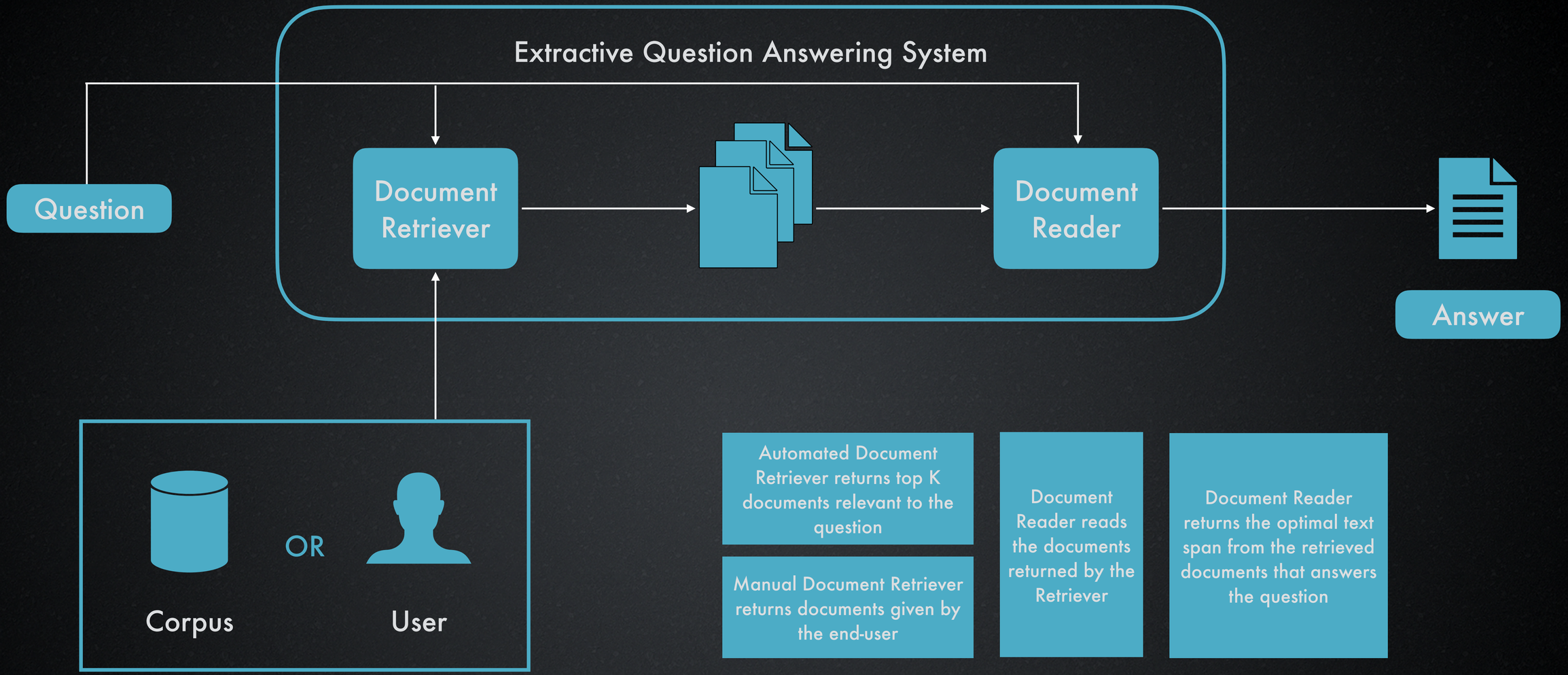
- Access to huge amounts of unlabeled privacy policy dump (recently published).
- Experimented with latest state-of the-art tuning strategies and best practices that showed remarkable improvement in performances.

Future work

- With access to more GPU compute power, we can tune and optimize the models even further.
- Experiment with more data, or better model architectures.

Question-Answering System

(QUERY Module)



Dataset
(PolicyQA)

- Used PolicyQA dataset curated from the OPP-115 corpus.
- Reading-comprehension style dataset that includes information about the annotated spans, corresponding policy segments, and the associated Practice, Attribute, Value triples derived from the OPP-115 corpus to form the examples in the dataset.
- 714 individual questions.

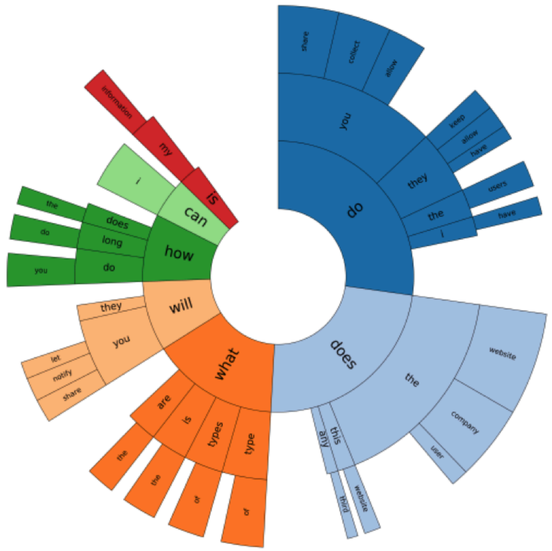


Figure 6: Distribution of trigram prefixes of questions in PolicyQA dataset.

Privacy Practice	Proportion	Example Question From PolicyQA
First Party Collection/Use	44.4 %	Why do you collect my data?
Third Party Sharing/Collection	34.1 %	Do they share my information with others?
Data Security	2.2 %	Do you use encryption to secure my data?
Data Retention	1.7 %	How long they will keep my data?
User Access, Edit and Deletion	3.1 %	Will you let me access and edit my data?
User Choice/Control	11.0 %	What use of information does the user choice apply to?
Policy Change	1.9 %	How does the website notify about policy changes?
International and Specific Audiences	1.5 %	What is the company's policy towards children?
Do Not Track	0.1 %	Do they honor the user's do not track preference?

Figure 7: OPP-115 categories of the questions in the PolicyQA dataset.

Source

Question-Answering System

(Approach)

- BERT tokenizer converts the input text into a model-ingestible format.
- Two linear classifiers predict the boundary of the evidence or the answer span
- When working with Question Answering, BERT requires the input chunk of text to be in specific format:

[CLS] question tokens [SEP] context tokens [SEP]

Vanilla BERT-base (uncased)

Vanilla RoBERTa-base (cased)

Domain-specific RoBERTa-base (cased)

Question-Answering System

(Training and Evaluation)

- PolicyQA comes in a similar setting as the Stanford Question Answering Dataset v2 (SQuAD2.0).
- Pre-train each of the three BERT models on the SQuAD2.0 dataset before fine-tuning on the domain-specific PolicyQA dataset.
- SQuAD2.0 consists of 150,000 answerable and unanswerable questions posed on a set of Wikipedia articles.
- Trained the models on SQuAD2.0 for two epochs using the HuggingFace's transformers library.
- Applied the BertAdam optimizer and further trained each model with a learning rate of 0.0003 and 20% dropout for five epochs on the PolicyQA dataset.

Evaluation Metrics:

Exact Match

F1-score

Comparing our results with previous benchmarks

(Question Answering)

PolicyQA results

BERT Size	Valid		Test	
	EM	F1	EM	F1
Tiny	21.0	47.1	15.5	39.9
Mini	26.5	55.2	22.8	49.8
Small	28.4	57.2	24.6	52.3
Medium	**31.1**	59.1	25.2	53.5
Base	30.5	**59.4**	**28.1**	**55.6**

Table 7: Performance of different sized QA models.

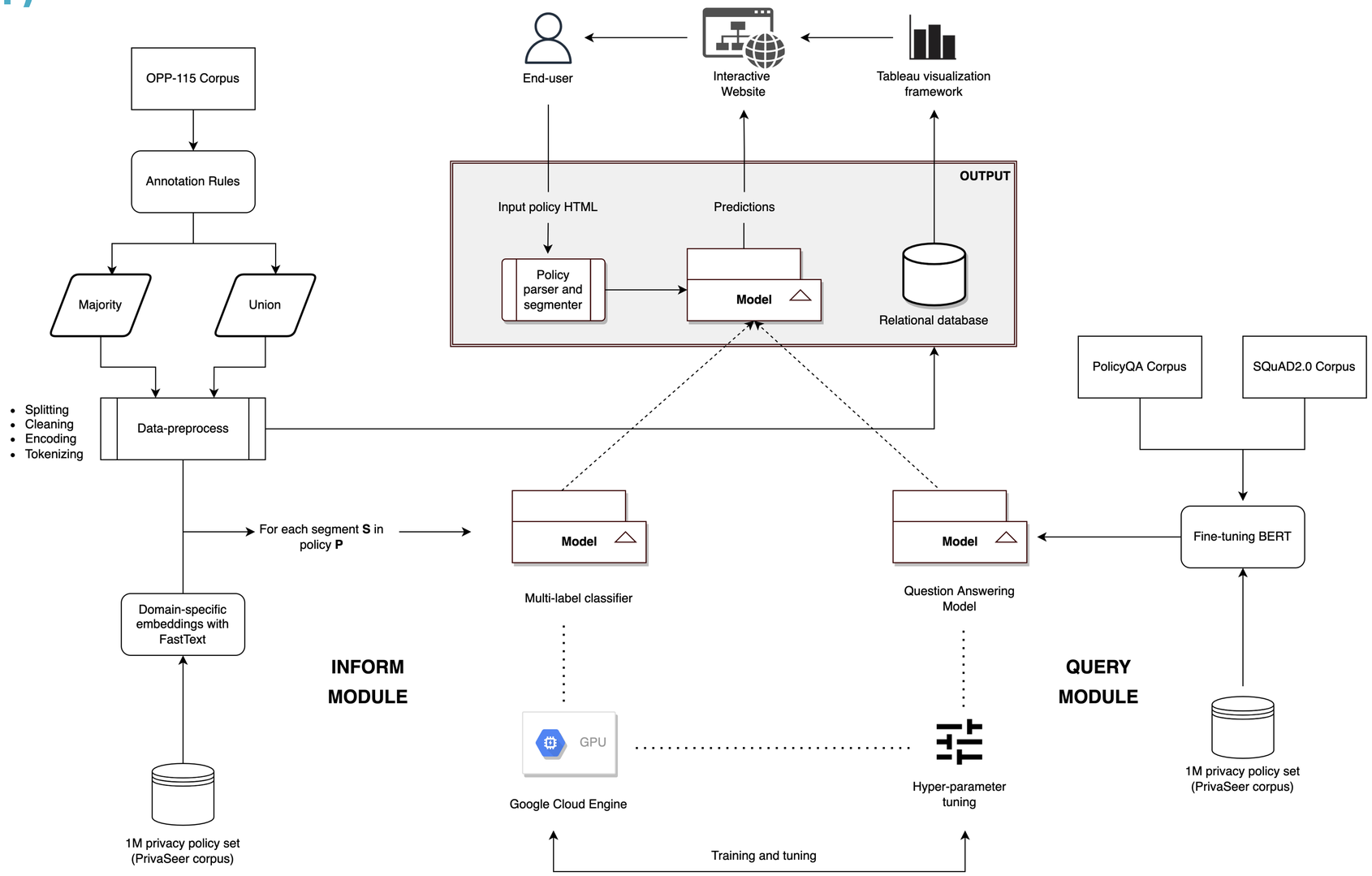
Our results

Model	**Valid**		**Test**	
	EM	**F1**	**EM**	**F1**
Vanilla BERT-base	30.61	59.70	28.25	56.08
Vanilla RoBERTa	32.92	61.51	31.02	58.42
Domain-specific RoBERTa	**35.57**	**63.16**	**32.20**	**59.48**

Table 3: Performance of our models on PolicyQA. The boldface values indicate the best performances.

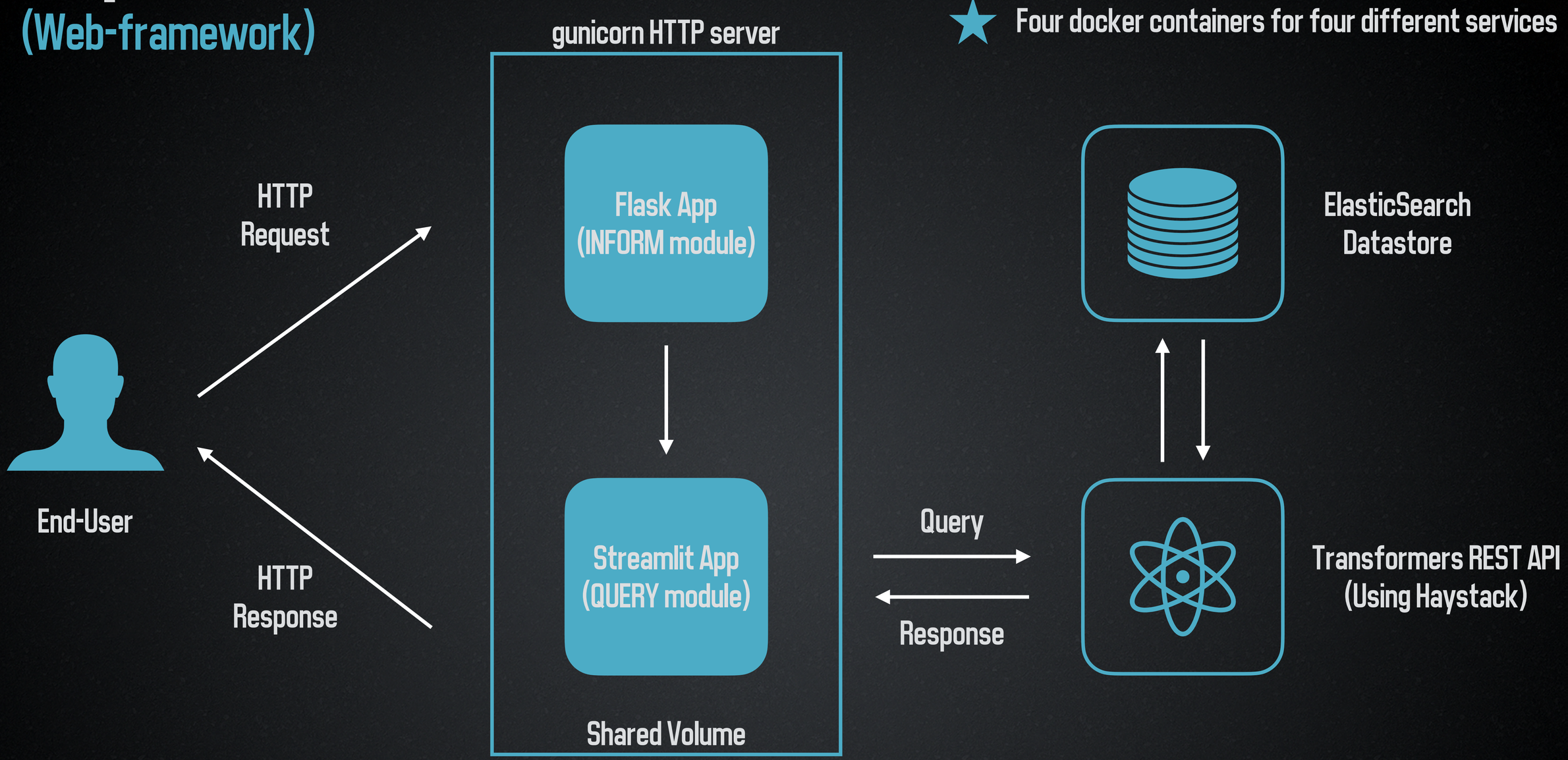
System Architecture

(INFORM + QUERY)



Implementation Design

(Web-framework)



Project Timeline

Full-semester project

- ☆ Literature Survey
- ☆ Preliminary Results

01 Phase 1 Pitch

- ☆ Data Preparation
- ☆ Visualizations
- ☆ Front-end (some progress)

02 Phase 1 Progress Updates

- ☆ Implement Front-end + Back-end (viz)
- ☆ Traditional models
- ☆ Language Neural Models (except BERT)

03 Phase 1 Presentation

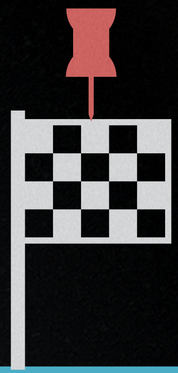
- ☆ Transformer models
- ☆ Literature Survey (QA systems)

04 Phase 2 Pitch

- ☆ QA using Transformers
- ☆ Front-end + Back-end

05 Phase 2 Progress Updates

06 Phase 2 Presentation



INFORM module — QUERY module

Key Takeaways

- The project was overall a success.
- Able to explore a fascinating problem of proliferating data privacy awareness to end-users using machine learning techniques.
- Our models produced promising results for both downstream tasks (Classification and Question Answering), outperforming previously published results on the same datasets.
- Faced some limitations due to computational and time constraints.
- Find ways to optimize the performance of our models by considering ensembling techniques and build more efficient and feasible language neural architectures.


THANK
YOU